Análise de Componentes Principais

TUANY DE PAULA CASTRO

Resumo

- **Objetivo:** reduzir uma grande massa de dados garantindo a menor perda possível de informação.

- **Metodologia:** novos eixos são encontrados a partir da rotação das variáveis originais garantindo o máximo da variabilidade explicada. Esses eixos são as componentes principais (combinações lineares das variáveis originais).

- **Propriedades:** as componentes principais são ordenadas decrescentemente de acordo com a variabilidade explicada e são independentes.

- **Duas construções:** a análise pode ser aplicada com a matriz de covariância ou correlação, sendo essa última mais indicada na maioria dos casos.

Resumo

- **Resultados:**
  - **Escolha das componentes:** critério de Kaiser, Scree Plot ou critério da variância acumulada.
  - **Contribuição das componentes:** a proporção da variância explicada por cada componente é igual a

$$\frac{autovalor\ da\ componente}{soma\ de\ todos\ os\ autovalores}$$

  - **Interpretação das componentes:** para interpretação, é necessário analisar as correlações entre as componentes escolhidas e as variáveis originais.

- **Observação:** A análise de Componentes Principais frequentemente serve de etapa intermediária em investigações maiores, como entradas numa regressão múltipla ou análise de agrupamento ou ainda como método de extração de fatores na análise fatorial.

Estudo da motivação e relação com a universidade

Foi aplicado a 107 estudantes um questionário com 28 questões para estudar os motivos que os levam à universidade. O objetivo da análise era identificar diferentes perfis de estudantes quanto à motivação do estudo e avaliar seu desempenho numa determinada disciplina, observando também se havia diferenças entre estudantes que assistiam e não assistiam vídeos motivacionais antes das aulas.

Para melhorar a assertividade do estudo, inicialmente aplicou-se a técnica de Componentes Principais com o intuito de reduzir a dimensão das 28 questões da bateria, explicando ao máximo a variabilidade dos dados.

Estudo da motivação e relação com a universidade

Etapa 1 – Análise de Componentes Principais

- Aplicou-se ACP nas 28 questões sobre a motivação de ir à universidade utilizando a matriz de correlação;
- Foram obtidas 6 componentes que explicaram 66,7% da variabilidade dos dados.
- Interpretação das componentes:
  - Componente 1 - Busca por excelência acadêmica
  - Componente 2 - Busca por uma boa carreira
  - Componente 3 - Prazer por frequentar a universidade
  - Componente 4 - Busca por uma boa remuneração futura
  - Componente 5 - Aumento da competência profissional
  - Componente 6 - Aprender coisas novas e ter boa remuneração

Estudo da motivação e relação com a universidade

Etapa 2 – Análise de Agrupamento

- A Análise de Agrupamento foi aplicada a fim de identificar os diferentes perfis de estudantes quanto à motivação em ir à universidade;
- Foram utilizadas como variáveis de entrada as 6 componentes principais obtidas com a ACP;
- Foram encontrados 3 grupos de alunos descritos na próxima tabela:

Estudo da motivação e relação com a universidade

Casela verde: média significativamente maior (menor do que 5%)

Casela vermelha: média significativamente menor (menor do que 5%)

Casela amarela: indício de diferença (de 5% a 10%)

	Total	1	2	3
Frequência Absoluta	**107**	26	71	10
Frequência Relativa	**100%**	24,30%	66,36%	9,35%
1. Busca por excelência acadêmica	**23,89**	29,21	22,62	19,09
2. Busca por uma boa carreira	**2,28**	1,83	2,96	-1,34
3. Prazer em frequentar a Universidade	**-5,80**	-6,71	-5,38	-6,47
4. Busca por uma boa remuneração no futuro	**4,74**	3,86	5,21	3,74
5. Aumento da competência profissional	**0,37**	-0,19	0,44	1,34
6. Aprender coisas novas e ter uma boa remuneração	**0,25**	0,31	0,04	1,60

Estudo da motivação e relação com a universidade

Nota-se que o Grupo 1 é formado por 26 alunos motivados pela busca da excelência acadêmica, sem muito prazer em frequentar a universidade e também sem muita preocupação com a carreira ou remuneração futura.

Já o Grupo 2, com 71 alunos, é mais voltado ao perfil de busca por uma boa carreira e boa remuneração futura. Trata-se do perfil mais comum nas universidades (maior grupo).

Por último, o Grupo 3 é formado pelos estudantes motivados pelo aumento da competência profissional e pouco interessados na busca por excelência acadêmica.

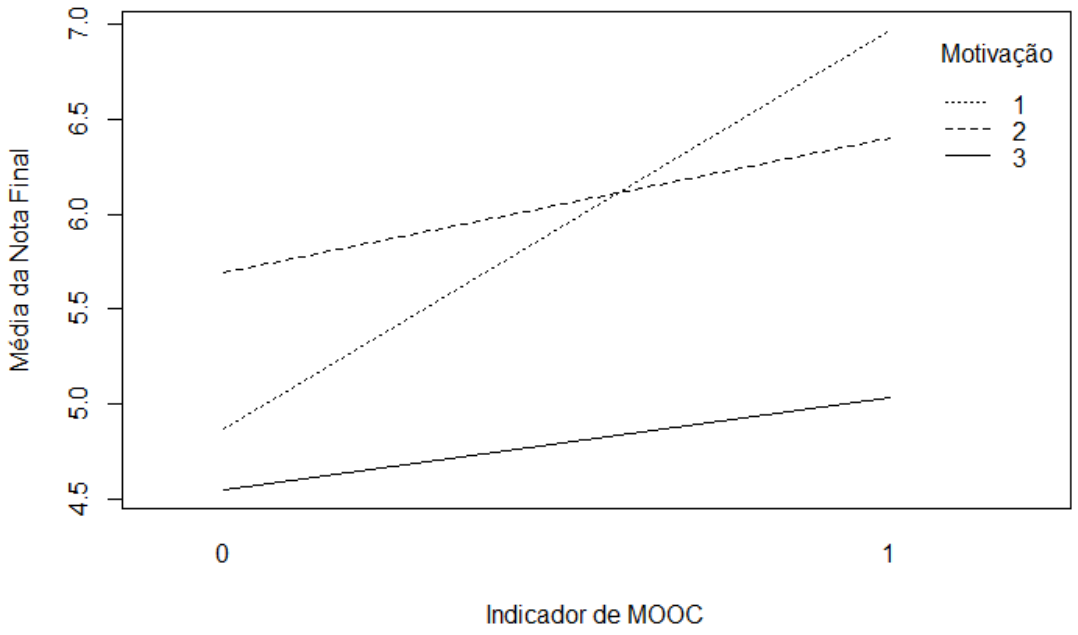
Estudo da motivação e relação com a universidade

Etapa 3 – Análise de Regressão

- Nessa última etapa, o objetivo da análise era avaliar se os vídeos foram importantes no desempenho final do aluno;
- Avaliou-se o efeito dos vídeos na nota final controlando por gênero (masculino ou feminino), turma de estudo (Gestão da Inovação, Liderança e Metodologia Científica) e grupo de motivação (1, 2 ou 3);
- Inicialmente, foi construído um gráfico de perfis da média das notas finais dos alunos de acordo com o grupo de motivação e o indicador de vídeo a fim de verificar efeito de interação entre as duas variáveis.

Estudo da motivação e relação com a universidade

Por meio do gráfico de perfis ao lado, pode-se observar que aparentemente há efeito de interação entre o grupo de motivação e o indicador de vídeo, pois a diferença de média final entre os que viram e não viram o vídeo não é igual nos três grupos de motivação, sendo maior no grupo 1.



Estudo da motivação e relação com a universidade

- De acordo com a observação do gráfico acima, foi construído um modelo de Regressão Normal para a nota final do aluno com variáveis explicativas:
  - Gênero (masculino ou feminino)
  - Turma (Gestão da Inovação, Liderança, Metodologia Científica)
  - Grupo de Motivação (1, 2, 3)
  - Indicador de vídeo (sim ou não)
  - Interação entre Grupo de Motivação e Indicador de vídeo

  Os resultados do modelo encontram-se na tabela seguinte:

Estudo da motivação e relação com a universidade

Variável	Estimativa	Limite Inferior	Limite Superior	Valor-p
Perfil base	6,0	5,0	7,1	< 0,1%
Gênero (Masculino)	-0,4	-1,0	0,1	12,2%
Turma (Liderança)	-1,8	-2,5	-1,1	< 0,1%
Turma (Metodologia científica)	0,3	-0,4	0,9	43,1%
Grupo de motivação (2)	0,4	-0,7	1,5	44,8%
Grupo de motivação (3)	-0,4	-1,9	1,0	55,1%
MOOC (Sim)	2,1	1,0	3,1	< 0,1%
MOOC (Sim):Grupo de motivação (2)	-1,3	-2,6	0,0	4,5%
MOOC (Sim):Grupo de motivação (3)	-1,3	-3,1	0,6	17,4%

*Perfil base: mulheres da turma de Gestão da Inovação, do primeiro grupo de motivação e que não assistiram ao MOOC.

Estudo da motivação e relação com a universidade

Nesse primeiro modelo, nota-se que não há diferença estatisticamente significativa em:

- Turmas de Metodologia Científica e Gestão da Inovação (valor-p = 43,1%);
- Gêneros masculino e feminino (valor-p = 12,2%);
- Efeitos de vídeo nos grupos 2 e 3 (podem ser considerados iguais).

Assim, construiu-se um novo modelo que segue na tabela seguinte:

Estudo da motivação e relação com a universidade

Variável	Estimativa	Limite Inferior	Limite Superior	Valor-p
Perfil base	5,9	5,0	6,8	< 1%
Turma (Liderança)	-1,8	-2,3	-1,3	< 1%
Grupo de motivação (2)	0,3	-0,7	1,4	55,1%
Grupo de motivação (3)	-0,5	-1,7	0,7	40,5%
MOOC (Sim)	2,2	1,2	3,2	< 1%
MOOC (Sim):Grupo de motivação (2-3)	-1,3	-2,5	-0,1	3,77%

Interpretações:

- **Perfil base:** A média final esperada dos alunos da turma de Gestão da Inovação ou Metodologia Científica, do primeiro grupo de motivação e que não assistiram ao MOOC é de 5,9 (5,0;6,8);
- **Turma (Liderança):** Alunos da turma de Liderança têm média final menor em 1,8 pontos, em média, do que os alunos das turmas de Gestão da Inovação ou Metodologia Científica, mantendo-se constantes as demais variáveis;

Estudo da motivação e relação com a universidade

- **MOOC (Sim) - Grupo de motivação (1):** Alunos do grupo 1 de motivação têm média final aumentada em 2,2 (1,2; 3,2) pontos, em média, quando assistem o MOOC, mantendo-se constante a Turma;
- **MOOC (Sim) - Grupo de motivação (2):** Alunos do grupo 2 de motivação têm média final aumentada em 1,2 (1,2; 3,2) pontos, em média, quando assistem o MOOC, mantendo-se constante a Turma;
- **MOOC (Sim) - Grupo de motivação (3):** Alunos do grupo 3 de motivação têm média final aumentada em 0,4 (-0,8; 1,6)* pontos, em média, quando assistem o MOOC, mantendo-se constante a Turma;

(*) Como o intervalo de confiança contém zero, o efeito pode ser considerado desprezível.

Exercícios

1) Em *dados-RH.xlsx,* encontram-se dados referentes a funcionários de uma empresa. As variáveis em estudo são:

- nível de satisfação (*satisfaction_level*),
- última avaliação (*last_evaluation*),
- número de projeto (*number_project*), horas mensais médias (*average_monthly_hours*),
- tempo gasto na empresa (*time_spend_company*),
- se eles tiveram um acidente de trabalho (*Work_accident*),
- se eles tiveram promoção nos últimos 5 anos (*promotion_last_5_years*),
- salário (*salary*),
- se o funcionário deixou a empresa (*left*)
- departamento de trabalho (*sales*).

a) Faça uma análise de componentes principais para resumir os dados e interprete.

Exercícios

b) Compare descritivamente funcionários e ex-funcionários com relação às componentes encontradas.

c) Faça um modelo de regressão logística para verificar como as componentes influenciam na chance de um funcionário deixar a empresa.

2) Em *dados-glass.xlsx*, observam-se os dados de classificação de tipos de vidro. Foram coletadas as seguintes variáveis:

- RI: índice de refração
- Na: percentual de sódio apresentado
- Mg: percentual de magnésio apresentado
- Al: percentual de alumínio apresentado

Exercícios

- Si: percentual de silício apresentado
- K: percentual de potássio apresentado
- Ca: percentual de cálcio apresentado
- Ba: percentual de bário apresentado
- Fe: percentual de ferro apresentado
- Type: 1 - janelas de construção processadas, 2 - janelas de construção não processadas, 3 - janelas de veículos processadas, 4 - janelas de veículos não processadas (não tem nesse banco de dados), 5 – contêineres, 6 – louças, 7 – faróis.

(a) Discuta os resultados da aplicação de uma análise de componentes principais para resumir as informações quantitativas da base de dados.

(b) Descreva as componentes principais encontradas.

Exercícios

c) Compare, descritivamente, os tipos de vidro com relação às componentes encontradas.

d) Utilize uma Análise Discriminante com as componentes encontradas e verifique a assertividade da classificação do tipo de vidro nos dados abaixo.

RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type
1,517	14,38	0	1,94	73,61	0	8,48	1,57	0	7
1,517	14,23	0	2,08	73,36	0	8,62	1,67	0	7
1,519	14	2,39	1,56	72,37	0	9,57	0	0	6
1,517	12,86	0	1,83	73,88	0,97	10,17	0	0	5
1,52	13,27	0	1,76	73,03	0,47	11,32	0	0	5
1,518	13,65	3,66	1,11	72,77	0,11	8,6	0	0	3
1,517	13,72	3,68	1,81	72,06	0,64	7,88	0	0	2
1,517	13,3	3,64	1,53	72,53	0,65	8,03	0	0,29	2
1,516	12,72	3,46	1,56	73,2	0,67	8,09	0	0,24	1
1,516	13,53	3,55	1,54	72,99	0,39	7,78	0	0	1

Exercícios

3) Os dados *Wine_data.csv* correspondem a 12 características avaliadas em 1599 vinhos.

a) Encontre e interprete as componentes principais que resumem as informações contidas nos dados.

b) Com os escores dessas componentes principais encontradas, faça uma análise de agrupamento para identificar e caracterizar os diferentes grupos de vinhos da amostra.

4) Em *food-comsuption.csv* encontram-se os dados referentes a hábitos de consumo alimentício em 16 países europeus.

a) Resuma as informações das 20 variáveis em componentes principais. Interprete.

Exercícios

b) Utilizando as componentes principais, compare por meio de medidas resumo os 16 países.

c) Faça uma análise de agrupamento e compare os grupos de países quanto às componentes em estudo.

5) Em r*aw-material-characterization*.csv, encontram-se 6 mensurações de características de lotes de pastilhas de plástico e ainda o resultado de avaliação do uso do material (fraco ou adequado). Algumas variáveis estão codificadas para confidencialidade.

- Número do lote (Lot number)
- Resultado da avaliação (Outcome)
- Percentuais de materiais 1, 2 e 3 num determinado intervalo de tamanho (Size5, Size10, Size15)
- Medida termogravimétrica (TGA);

Exercícios

- Medida termomecânica (TMA);
- Medida de calorimetria (DSC).

a) Faça uma análise de Componentes Principais para resumir as informações contidas nas variáveis, exceto *Outcome*. Interprete os resultados.

b) Compare descritivamente as pastilhas fracas e adequadas quanto às componentes encontradas.

c) Faça um modelo de regressão logística para verificar a significância das componentes na chance de uma pastilha ser adequada.

d) Quais seriam suas recomendações para se obter mais lotes considerados adequados do que fracos?

Referências

- BISQUERRA, R; CASTELLA, J.; VILLEGAS, F. Introdução à estatística: enfoque informático com o pacote estatístico SPSS. Porto Alegre: Artmed, 2007.
- DANCEY, Cristine P; REIDY, John. Estatística sem Matemática Para Psicologia. 3 edição. Porto Alegre: Artmed, 2006.
- HAIR, J.; ANDERSON, R.; BLACK, W. Análise multivariada de dados. 6 ed. Reimp. Porto Alegre: Bookman, 2009.
- JOHNSON, R. and WICHERN, D. Applied Multivariate Statistical Analysis. Sixth edition, Wisconsin, Pearson, 2007.